基于SVM的多Agent信息融合算法*)

马 骏 张健沛 杨 静 程丽丽

(哈尔滨工程大学计算机科学与技术学院 哈尔滨 150001)

摘 要 支持向量机(SVM)是一种基于结构风险最小化原理、具有很高泛化性能的学习算法,为小样本、非线性、高维数一类信息融合问题的建模提供了一种有效的途径。本文将Mobile Agent运用到信息融合系统中,对信息融合系统中原有OODA模型进行改进,提出了一种基于SVM的Mobile Agent信息融合模型及算法。相关实验表明,本文中的训练算法可达到更为满意的分类效果,并且可以得到较高的分类精度。

关键词 支持向量机,移动Agent,信息融合

An Algorithm of Multi-Agent Information Fusion Based on SVM

MA Jun ZHANG Jian-Pei YANG Jing CHENG Li-Li

(College of Computer Science and Technology, Harbin Engineering University of China, Harbin 150001)

Abstract The support vector machine (SVM) is an algorithm based on structure risk minimizing principle and has high generalization ability. The model offers a kind of effective way for the information fusion problem of little sample, non-linear and high dimension. In this paper, mobile agent is applied to information fusion system. The model of OODA and the study method of information fusion system are improved. The model and an algorithm of information fusion based on the support vector machine are proposed. The experiment results show that this hierarchical and parallel SVM training algorithm is efficient to deal with large-scale classification problems and has more satisfying accuracy in classification precision.

Keywords Support vector machine(SVM), Mobile agent, Information fusion

1 引言

信息融合技术[1]是近20年来随着微电子技术和计算机技术的发展而发展起来的一门新技术。目前关于信息融合概念的描述可概括为:将某一目标的多源信息进行融合,形成比单一信息源更精确、更完整的估计和判决,也就是把在空间或时间上冗余或互补的数据,依据某种准则进行组合,以获得被测对象的一致性描述或理解。

近年来,许多专家和学者提出了多种信息融合模型,其共同点或中心思想是在信息融合过程中进行多级处理。现有系统模型大致可以分为两大类:根据节点顺序构建的功能型模型,如UK情报环、Boyd控制回路OODA环[2];根据数据提取加以构建的数据型模型,如JDL模型、瀑布模型等。

信息融合技术有很多方法,其中比较成熟的有卡尔曼滤波法、贝叶斯推理方法和D-S证据推理方法等。但是,在实际运用中,这些方法都有各自的问题,很难获得较好的处理效果。它们中有的依赖于先验概率分布及条件概率分布;有的依赖于信度函数及模糊隶属度。另外,对于一类小样本、高维特征空间条件下的信息融合问题,迄今仍没有一个有效的方法。而SVM具有更严格的理论和数学基础,不存在局部最小问题。小样本学习使它具有很强的泛化能力,不过分依赖样本的数量和质量。本文将SVM理论引入到信息融合研究中,在改进现有模型的基础上,提出了一种基于SVM理论的信息融合算法,该算法较好地解决了高维特征空间和不确定条件下的信息融合问题。

2 支持向量机

支持向量机[6]建立在统计学习理论的VC维理论和结构风险最小化原理的基础上,根据有限的样本信息,在模型的复杂性(对特定训练样本的学习精度)和学习能力(无错误地识别任意样本的能力)之间寻求最佳折中,以获得较好的综合能力。该方法已在文本、图像识别与分类中得到了较好的应用和研究。

在线性可分的情况下,假设两类样本,$(x_1,y_1),\cdots,(x_l,y_l)$,$x\in R^n$,$l$为样本数,$n$为输入维数,有一个超平面将这两类完全分开。这样的超平面可描述为

$$(w\cdot x)+b=0 \tag{1}$$

如果样本被无误差地划分且超平面到两类的距离最大,则称这个超平面为最优超平面。

求解最优超平面可以看成求解二次型规划的问题。对于训练样本集合S,找到权值w和偏移b的最优值,使权值代价函数最小:

$$\min \Phi(w)=\frac{1}{2}\|w\|^2 \tag{2}$$

约束满足条件:

$$y_i(w\cdot x_i+b)-1\geqslant 0, i=1,2,\cdots,l \tag{3}$$

优化函数$\Phi(w)$为二次型,约束条件是线性的,因此是典型的二次规划问题,可由Lagrange方法求解。引入Lagrange乘子,$a_i\geqslant 0, i=1,2,\cdots,l$:

$$L(w,b,a)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{l}a_i\{y_i(x_i\cdot w+b)-1\}L$$

的极值点为鞍点,取L对w和b的最小值$w=w^*$,$b=b^*$,以及对a的最大值$a=a^*$。对L求导,通过求解二次规划,解得对应的a^*和w^*:

$$w^*=\sum_{i=1}^{l}a_i^*y_ix_i \tag{4}$$

以及最优超平面。那些$a_i=0$的样本对于分类问题不起什么

*)本课题得到国家自然科学基金(60673131)和黑龙江省自然科学基金(F2005-02)。马 骏 博士生,主要研究方向:数据挖掘。

作用，只有 $a_i=0$ 的样本对 w^* 起作用，并决定分类结果。这样的样本被定义为支持向量。

a^* 和 w^* 可以显式求得，选取一个支持向量样本 x_i：

$$b^*=y_i-w\cdot x_i \tag{5}$$

对于任一输入样本 x，计算分类函数：

$$d(x)=x\cdot w^*+b^*=\sum_{i=1}^{l}y_i a_i^*(x\cdot x_i)+b^* \tag{6}$$

分类函数 $d(x)$ 的符号确定样本 x 的归属。如果样本是线性不可分，则可以通过核函数定义的非线性变换，将线性不可分问题转化为线性可分问题。

3 信息融合系统模型及结构

OODA 环在信息融合系统中极具代表性，包括四个处理阶段：决策、定向、观测和执行。在许多信息融合系统中，对 OO(观测和定向)阶段进行了大量的研究，主要是对环境的态势信息的获取以及融合方法的选择[8]。

Mobile Agent 技术可有效地简化分布式系统的设计、实现和维护。Mobile Agent 是独立运行的计算机程序，代表用户完成特定的任务，具有自主性、移动性、协作性、安全性和智能性等特性。MA 计算模式能有效地降低分布式计算中的网络负载，提高通信效率，支持异步及自主交互，支持非连接互操作，可动态自适应，具有移动的坚定性和容错能力。

利用 MA 优越的性能，在 OODA 模型的基础上，采用 MA 获取环境信息[3]。可以把对环境的观测由 MA 来完成，观测的结果送到融合中心进行定向，然后由决策中心根据融合数据对环境做出判断，通过执行使 MA 按照决策更有效地获取环境信息。改进的模型如图 1。

关于信息融合系统的结构目前尚未形成统一的分类形式，但大体上可以把融合系统的结构分为三类：集中式、分散式以及分级式[7]。集中式融合结构的融合速度快，各融合中心计算和通信负担过重，系统容错性差，各低层传感器之间缺乏必要的联系。分散式融合结构中，每个传感器都具有估计全局信息的能力，任何一个传感器的失效都不会导致系统的崩溃，但系统的通讯费用很高。分级融合结构各传感器之间是一种层间有限联系，其计算和通信负担介于集中式结构和分散式结构之间。分级融合信息从低层向高层逐层流动，无反馈时，层间传感器属于单向联系，高层信息不参与低层处理；有反馈时，层间传感器是双向联系，不仅低层融合信息向高层传递，高层信息也参与低层节点处理。具有反馈的分级融合结构如图 2 所示。

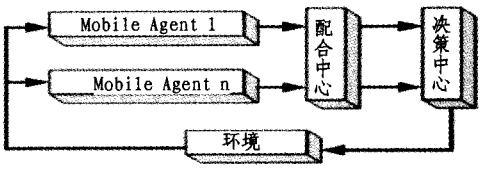

图 1 基于 Mobile Agent 的 OODA 环模型

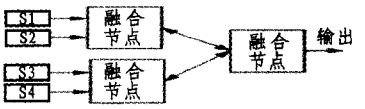

图 2 具有反馈的分级融合结构

4 基于 SVM 的信息融合算法

多 Agent 信息融合是指对来自多个 Mobile Agent 信息源的数据进行检测、关联、相关、估计和综合等多级、多方面的处理，以获得对被测对象状态的精确估计和评价。利用多个 Mobile Agent 信息源所获取的关于对象和环境的信息，根据不同任务所需要的全面、完整的信息，主要体现在融合算法上。因此，信息融合的核心问题是选择合适的信息融合算法。

4.1 基于 SVM 的信息融合结构

本文在融合节点上采用 SVM 方法[5]，并对原有反馈分级结构进行改进，采用交叉反馈的方式来进行信息融合[4]。交叉反馈的并行方式能有效地缩短 SVM 训练时间，具有良好的可扩充性。将 SVM1，SVM2 和 SVM5 三个分类器的组合作为第一个融合节点，SVM3，SVM4 和 SVM6 的分类器组合作为第二个融合节点，将两个融合节点的融合结果发送到第三个融合节(SVM7)。基于 SVM 的信息融合结构如图 3 所示。

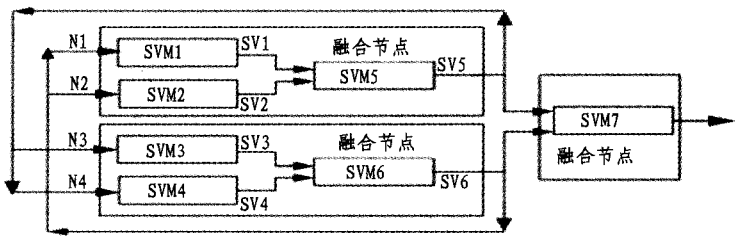

图 3 基于 SVM 的信息融合结构

假定训练数据子集 $N1,N2,N3,N4$ 相互独立。在图 3 结构中，两个 SVM 分类器对应的支持向量集被合并，如此继续，直至得到一个最后的支持向量集。其优势在于每一个 SVM 不需要处理所有训练数据，多个 SVM 分类器可以通过分布式系统训练得到，因此整体的训练时间被大大减少[9]。通常地，这种分级结构只需从头到尾运行一次就可以得到较为满意的结果，但是要达到全局的优化，还应在分类过程中加入反馈。

考虑训练数据集的不同分布状态和全局优化的需要，反馈是需要被用于调节多分类器的性能。我们对分级并行结构进行改进：将 $SV5,SV6$ 作为反馈加入到第一层数据中，进行交叉，$SV5$ 加到 $N3,N4$ 中，$SV6$ 加到 $N1,N2$ 中，这样避免数据的重复分类。

4.2 算法描述

具体算法如下：

1. 对由 MA 获取的数据子集 $N1,N2,N3,N4$ 进行训练，获得相应的 SVM 分类器和与之对应的支持向量集 $SV1$，$SV2,SV3,SV4$；

2. 将所得到的支持向量集两两合并，得到下一层训练数据集；

3. 训练 $SV1,SV2,SV3,SV4$，得到相应的 SVM 分类器 SVM5 和 SVM6，以及支持向量集 $SV5,SV6$；

4. 将 $SV5,SV6$ 作为反馈加入到第一层数据中，$SV5$ 加到 $N3,N4$ 中，$SV6$ 加到 $N1,N2$ 中；

5. 重复 1～3 步骤，直到获得新的 $SV5$，$SV6$；

6. 将通过反馈得到的新的 $SV5$，$SV6$ 合并，得到最后的输出 $SV7$ 和 SVM 决策函数；

7. 如果 $SV5$($SV6$)与通过反馈得到的 $SV5$($SV6$)相同，或者其差集中元素固定，则算法结束；

8. 如果不满足 7 中的条件，则返回第 5 步，重新开始训练。

5 实验及结果分析

先将现实中需要分类的数据集平均分成 4 个训练子集，分别训练得到 4 个 SVM 分类器，用事先人工分类的 4235 个数据样本作为测试集，生成的 78963 个样本为数据集，其维数为 26。

所有的程序使用 MATLAB7.0 编写，通过使用 MATLAB SVM Toolbox 训练仿真比较。采用标准的 SVM 算法以及径向基函数，即

$$K(x,y)=\exp\left[-\frac{(x-y)^2}{2\delta^2}\right]$$

为了说明分级并行算法的性能，我们在同一数据集上进行了三次实验。

实验 1 为标准 SVM，实验 2 为分级并行 SVM(无反馈)，实验 3 为分级并行 SVM(有反馈)。实验 1 和实验 2 使用传统的方法选择支持向量，实验 3 中使用本文改进的分级方法。表 1 列出了这三个实验的比较结果，最终输出的是支持向量的个数和分类精度。

表 1 实验结果

	E1	E2	E3
Number of SVs	2025	2356	1935
Output precision(%)	94.46	92.07	96.18
Training time(s)	30	18	25.2
feedback	no	no	yes

实验结果分析：实验 3 中，训练时间要比标准 SVM 算法少，具有较高的分类精度。实验 2 中，具有最少训练时间，因其忽略了所有的反馈，但分类精度最低，这个结果符合文中的分析与预测。实验 1 中，无论是分类精度，还是训练时间都无法达到让人满意的程度。

结论和进一步研究 本文将信息融合技术与 Mobile Agent 相结合，改进了 OODA 模型，提出了一种基于 SVM 的分级信息融合算法。该算法以分级结构为基础，分别由多个 SVM 分类器进行并行训练。在信息融合过程中，采用支持向量机的分级学习算法是可行的，可有效解决小样本、非线性参数之间存在模糊关系的信息融合问题。反馈和支持向量的选择是算法对标准反馈分级结构的改进，由于各局部优化独立解决，整体的存储空间和计算时间的消耗被大大降低。实验结果表明，本文中的训练算法可达到更为满意的分类效果，并可以得到较高的分类精度。此外，核函数的类型及相关参数的选择对融合精度有一定影响，如何优化，我们将对这些问题进行进一步的研究。

参 考 文 献

1 Varshney P K. Distributed Detection and Data Fusion. New York: Springer- Verlag, 1996

2 Shahbazian E, Dale E, Blodgett P L. The Extended OODA Model for Data Fusion Systems. In: Proceeding of International Conference on Information Fusion, USA, 2001

3 Shaban K B. Information Fusion in a Cooperative Multi-agent System for Web Information Retrieval. In: Proceedings of the Fifth International Conference on Information Fusion, Singapore, 2002

4 Kim Yongseog. Information Fusion via a Hierarchical Neural Network model. Journal of Computer Information Systems, 2005(4): 1～13

5 田盛丰，黄厚宽. 基于支持向量机的数据库学习算法. 计算机研究与发展，2000，37(1)：7～22

6 Vapnik V N. The Nature of Statistical Learing Theory. New York: Springer- Verlag, 1995

7 Kumar R, Wolenetz M, Agarwalla B. A Framework for Distributed Data Fusion. Information Fusion, 2007, 8(3): 227～251

8 Andler S F, Niklasson L, Persson O B. A Information Fusion from Databases, Sensors and Simulations: a Collaborative Research Program. In: 29th Annual IEEE/NASA Software Engineering Workshop, 2006

9 Wen Y M, Lu B L. A cascade method for reducing training time and the number of support vectors. In: Proceeding of International Symposium on Neural Network, Dalian, 2004

(上接第 160 页)

结论 本文针对 Bagging 和 Boosting 的不足，提出了一种新的算法 CG-Bagging。该算法能够在小规模集成的基础上得到较好的泛化能力，并且避免类似于 Boosting 产生的过度拟合的不足。在 UCI 机器学习数据库上的实验结果表明，CG-Bagging 的泛化能力略强于 Decorate 和 RandomForest，但其效率远优于 Bagging 和 Boosting。进一步的工作包括如何从其他途径生成平均泛化误差小且个体差异度大的分类模型，以提高集成的泛化能力。

参 考 文 献

1 Dietterich T G. Machine learning research: Four current directions. AI Magazine, 1997, 18(4): 97～136

2 Krogh A, Vedelsby J. Neural network ensembles, cross validation and active learning. In: G Tesauro, D S Touretzky and T K Leen, eds. Advances in Neural Information Processing Systems 7. Cambridge, MA: MIT Press, 1995

3 Breiman L. Bagging predicators. Machine Learning, 1996, 24(2): 123～140

4 Schapire R E. The boosting approach to machine learning: An overview. In: MSRI Workshop on Nonlinear Estimation and Classification, 2001

5 Melville P, Mooney RJ. Constructing Diverse Classifier Ensembles using Artificial Training Examples. In: Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence. Acapulco, Mexico, 2003. 505～510

6 Breiman L. Random Forests. Machine Learning, 2001, 45(1): 5～32

7 Zhou Z H, Wu J, Tang W. Ensembling Neural Networks: Many Could Be Better Than All Artificial Intelligence, 2002, 137: 239～263

8 Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques, 2nd edition. Morgan Kaufmann, San Francisco, 2005

9 Freund Y, Schapire R E. Experiments with a new boosting algorithm. In: Proceedings of the Thirteenth International Conference on Machine Learning. San Francisco, 1996. 148～156

10 Blake C L, Merz C J. UCI Repository of Machine Learning Databases. Available at: http://www.ics.uci.edu/～mlearn/Mlrepository.html, 1998

11 Quinlan J R. Bagging, Boosting, and C4.5. In: Proceedings of the Thirteenth National Conference on Artificial Intelligence. Cambridge, MA: AAAI Press/MIT Press, 1996

12 Quinlan J R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993